
Nonparametric Bayesian Density Modeling with Gaussian Processes

Ryan Prescott Adams¹

Iain Murray²

David J.C. MacKay¹

RPA23@CAM.AC.UK

MURRAY@CS.TORONTO.EDU

MACKAY@MRAO.CAM.AC.UK

¹Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, UK

²Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3G4, CA

Introduction

The Gaussian process is a useful prior on functions for Bayesian kernel regression and classification. Density estimation with a Gaussian process prior is difficult, however, as densities must be nonnegative and integrate to unity. The statistics community has explored the use of a logistic Gaussian process for density estimation, relying on approximations of the normalization constant (e.g. [1, 2, 3]).

We propose the Gaussian Process Density Sampler (GPDS), a nonparametric, practical and consistent method of constructing a Markov chain on the posterior distribution of an unknown density, without approximation. To our knowledge this is the first fully Bayesian approach to GP-based density estimation that does not require approximation of the normalization constant.

The Prior on Density Functions

We are concerned with densities on a space \mathcal{X} that permits a positive definite kernel. We define a prior distribution on densities over \mathcal{X} via a Gaussian process prior over functions $g(x) : \mathcal{X} \rightarrow \mathbb{R}$ so that each g corresponds to a density f via

$$f(x) = \frac{1}{\mathcal{Z}_\pi[g]} \Phi(g(x)) \pi(x) \quad (1)$$

where $x \in \mathcal{X}$, $\pi(x)$ is an arbitrary probability measure on \mathcal{X} , $\Phi(\cdot)$ is a nonnegative function with a finite upper bound c , and $\mathcal{Z}_\pi[g]$ is the normalization constant.

We generate exact and exchangeable data samples from a common density drawn from this prior by a rejection sampling procedure, shown in Algorithm 1, that “discovers” the sample of g as it proceeds. The algorithm maintains a *conditioning set* that is composed of all previously sampled inputs \mathbf{X} and outputs \mathbf{G} .

Samples generated via this procedure are *exact* in the sense that they are not biased by the starting state of a finite Markov chain. They are also *exchangeable*,

i.e. the probability of a given set of data is invariant under reordering. Note also that this procedure does not require evaluation of the function at more than a finite number of locations and does not require evaluation of the normalization constant $\mathcal{Z}_\pi[g]$. Examples of samples from this prior are shown in Figure 1.

Algorithm 1 Generate P samples from the prior

Purpose: Draw P exact samples from a common density on \mathcal{X} drawn from the prior in Equation 1

Inputs: GP hyperparameters θ

- 1: Initialize empty conditioning sets for the Gaussian process: $\mathbf{X} = \emptyset$ and $\mathbf{G} = \emptyset$
 - 2: **repeat**
 - 3: Draw a proposal from the base measure: $\tilde{x} \sim \pi(x)$
 - 4: Sample the function from the Gaussian process at \tilde{x} : $\tilde{g} \sim \mathcal{GP}(g | \mathbf{X}, \mathbf{G}, \tilde{x}, \theta)$
 - 5: Draw a uniform variate on $[0, c]$: $r \sim \mathcal{U}(0, c)$
 - 6: **if** $r < \Phi(\tilde{g})$ (Acceptance rule) **then**
 - 7: Accept \tilde{x}
 - 8: **else**
 - 9: Reject \tilde{x}
 - 10: **end if**
 - 11: Add \tilde{x} and \tilde{g} to the conditioning sets: $\mathbf{X} = \mathbf{X} \cup \tilde{x}$ and $\mathbf{G} = \mathbf{G} \cup \tilde{g}$
 - 12: **until** P samples have been accepted
-

Inference

We have N data $\mathcal{D} = \{x_n\}_{n=1}^N$ which we model as having been drawn independently from an unknown density $f(x)$. We use the prior from the previous section to specify our beliefs about $f(x)$, and we wish to generate samples from the posterior distribution over the latent function $g(x)$ corresponding to the unknown density. We may also wish to generate samples from the predictive distribution and perform hierarchical inference of the prior hyperparameters.

By using this prior to model the data, we are asserting that the data can be explained as the result of the procedure described in the previous section. We do not, however, know what rejections were made en route to accepting the observed data. These rejections are crit-

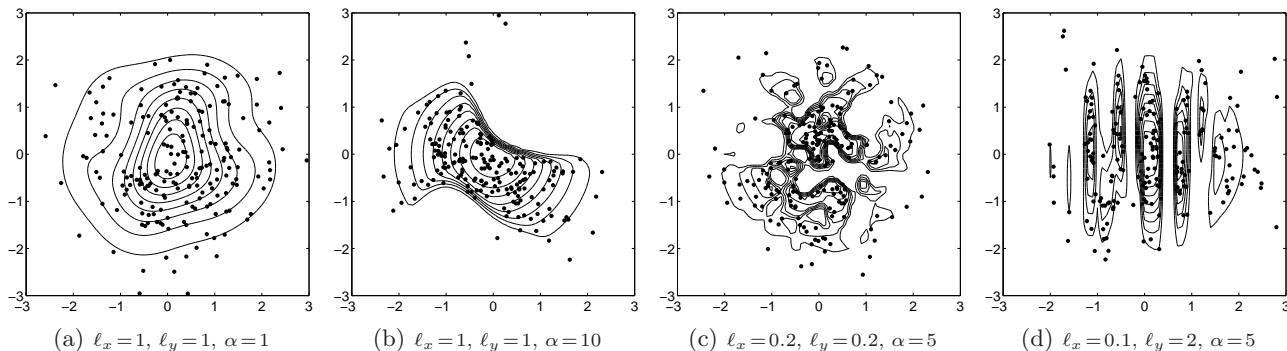


Figure 1. Four samples from the density prior are shown, along with 200 data samples. The contour lines show the approximate unnormalized densities. In each case the base measure is the zero-mean spherical Gaussian with unit variance. The covariance function was the squared exponential: $K(x, x') = \alpha \exp(-\frac{1}{2} \sum_i \ell_i^{-2} (x_i - x'_i)^2)$, with parameters varied as labeled in each subplot.

ical to defining the latent function $g(x)$. One might think of defining a density as analogous to putting up a tent: nailing down stakes is just as important as putting up poles. In density modeling, defining regions with little probability mass is just as important as defining the areas with significant mass.

While the rejections are not known, the generative procedure provides a probabilistic model that allows us to traverse the posterior distribution over possible *latent histories* that resulted in the data. This approach to inference is similar in spirit to that discussed by Murray [4], however we apply it to rejection sampling rather than coupling from the past [5]. We define a Markov chain whose equilibrium distribution is the posterior distribution over latent histories, and simulate plausible explanations of every step taken to arrive at the data. Such samples capture all the information available about the unknown density, and with them we may ask additional questions about $g(x)$ or draw predictive samples.

We model the data as having been generated exactly as in Algorithm 1, with $P = N$, i.e. run until exactly N proposals were accepted. The state space of the Markov chain on latent histories in the GPDS consists of: 1) the values of the latent function $g(x)$ at the data, 2) the number of rejections M , 3) the locations of the M rejected proposals, and 4) the values of the latent function $g(x)$ at the M rejected proposals.

We perform Metropolis–Hastings in two stages. First, we propose modifying the number of latent rejections M . Second, we propose moving the locations of the rejections and adjusting the function values, given M . We optionally add a third step that proposes new hyperparameters, conditioned on the entire latent history. To generate samples from the predictive distribution, after each M–H step we simply run the generative procedure forward from the current latent history.

Discussion

Computationally, there are at least two issues with the GPDS. First, the expected number of rejections for a given $g(x)$ is $N(\mathcal{Z}_\pi[g]^{-1} - 1)$. In high dimensional problems with a great deal of structure, we expect $\mathcal{Z}_\pi[g]$ to be small and so there would be many rejections. Second, the Gaussian process itself is expensive. The cost for taking a Metropolis–Hastings step is $O((M + N)^3)$, as it is necessary to decompose a matrix to evaluate the relative probabilities of two latent histories.

Valid MCMC algorithms for fully Bayesian kernel regression methods are well-established. This work introduces the first such method for density estimation, complementing alternatives such as Dirichlet Diffusion Trees [6] and infinite mixture models.

REFERENCES

- [1] Tom Leonard. Density estimation, stochastic processes and prior information. *Journal of the Royal Statistical Society, Series B*, 40(2):113–146, 1978.
- [2] Peter J. Lenk. Towards a practicable Bayesian nonparametric density estimator. *Biometrika*, 78(3):531–543, 1991.
- [3] Surya T. Tokdar and Jayanta K. Ghosh. Posterior consistency of logistic Gaussian process priors in density estimation. *Journal of Statistical Planning and Inference*, 137:34–42, 2007.
- [4] Iain Murray. *Advances in Markov chain Monte Carlo methods*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, London, 2007.
- [5] James G. Propp and David B. Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9(1&2):223–252, 1996.
- [6] Radford M. Neal. Defining priors for distributions using Dirichlet diffusion trees. Technical Report 0104, Dept. of Statistics, University of Toronto, 2001.